Abalone Age Prediction

Jieying Ma - vjyma@ucdavis.edu
Hebi Wang - hbwwang@ucdavis.edu

Abstract: A prediction model of abalone age is built on a dataset containing 9 variables of 4177 abalones. There is a useful model to predict the age of abalone by using the weight, height, diameter and infant factor.

1 Introduction

1.1 Background

The age of abalone is usually estimated by observing growth rings, a costly process that requires specialized skills and equipment. Therefore a method of predicting age by other easily accessible characteristics of abalone is needed. This project will be conducted on a dataset containing 9 variables of 4177 abalones, which are sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, and rings.[1]

1.2 Questions of Interest

1. How abalone age, which can be obtained from Rings+1.5, can be predicted by other variables.
2. Whether there are relationships among the whole weight, shucked weight, viscera weight, and shell weight of abalones.
3. Whether there are distinct trends in other abalone variables based on their gender.

1.3 Motivation

Exploring the relationship between abalone age and other variables can contribute to fisheries management. Divers can determine whether or not to catch an abalone by observing its distinctive features. This can prevent overfishing and maintain abalone abundance.

2 Methods and Results

2.1 Data Description

First, we use pie charts and histograms to visualize the distribution of the data.

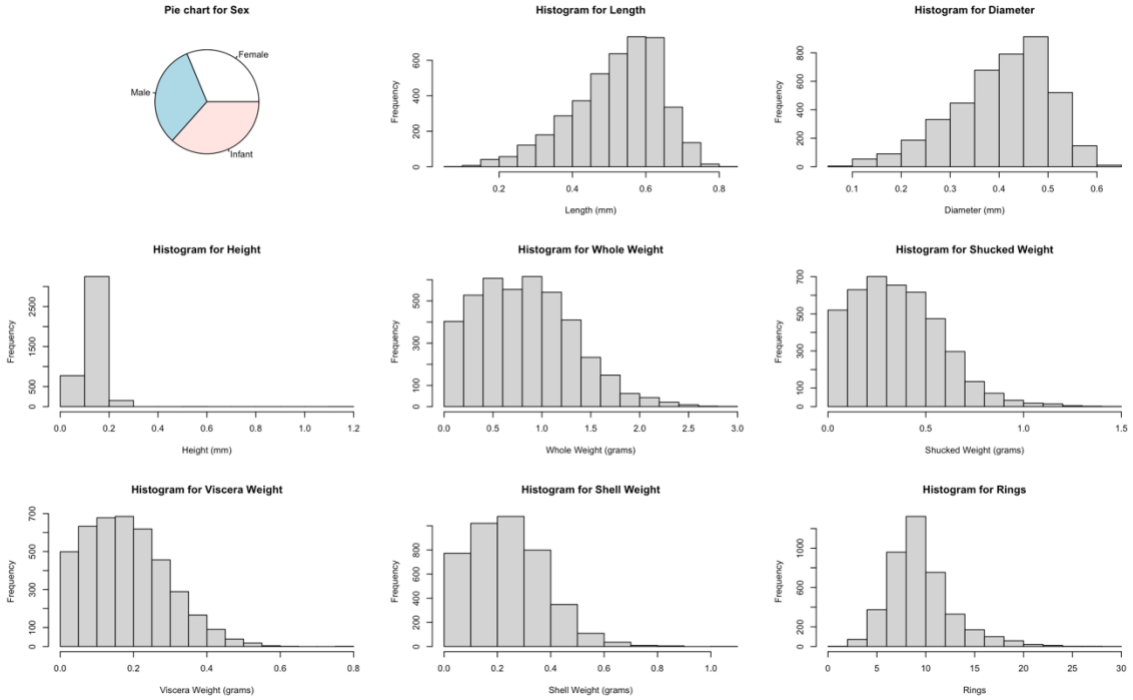

Figure 1. Variables distribution

From Figure 1 we can see that the response variable Rings is severely right-skewed, so it needs transformation. Since the "Height" has outliers greater than 0.4, we removed the cases where the "Height" is greater than 0.4. Then we perform the statistical analysis of the data without outliers.

Table 1. Summary of descriptive statistics for numerical variables

Variables	Mean	SD	Median	MAD	Min	Max	SE
Length (mm)	0.524	0.120	0.545	0.119	0.075	0.815	0.002
Diameter (mm)	0.408	0.099	0.425	0.096	0.055	0.650	0.002
Height (mm)	0.139	0.038	0.140	0.037	0.000	0.250	0.001
Whole weight (g)	0.828	0.490	0.800	0.529	0.002	2.826	0.008
Shucked weight (g)	0.359	0.222	0.336	0.235	0.001	1.488	0.003
Viscera weight (g)	0.181	0.110	0.171	0.118	0.001	0.760	0.002
Shell weight (g)	0.239	0.139	0.234	0.148	0.002	1.005	0.002
Rings	9.934	3.225	9.000	2.965	1.000	29.000	0.050

Since the values and ranges of all variables except rings are essentially within 2, no data normalization is needed.

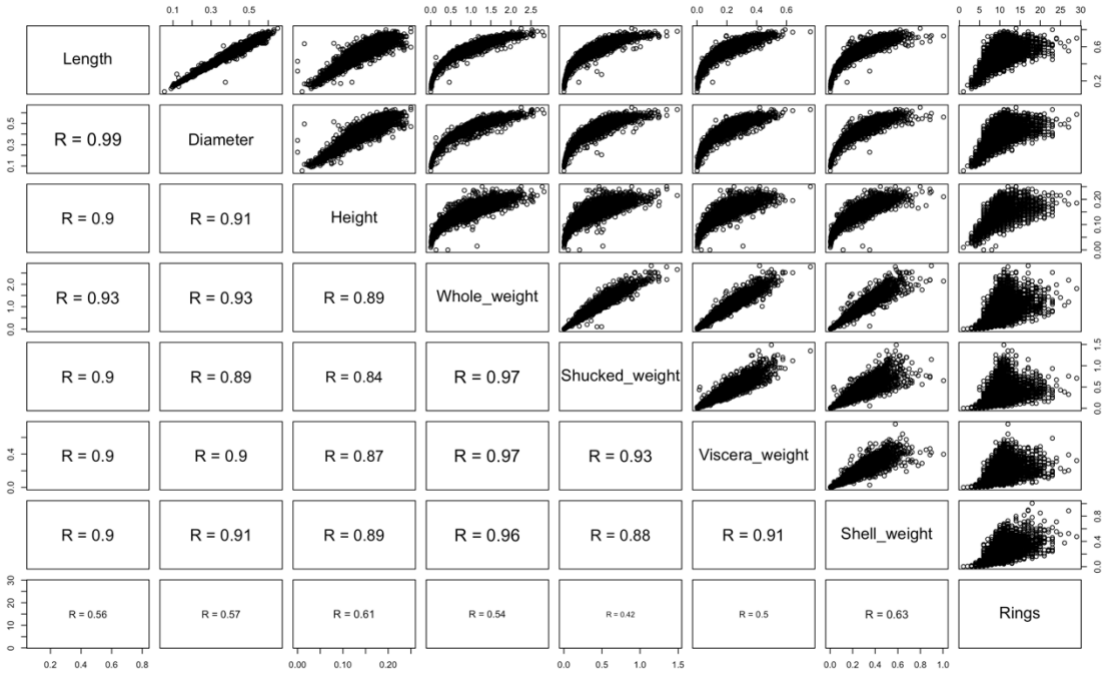

Figure 2. Numerical variables correlation

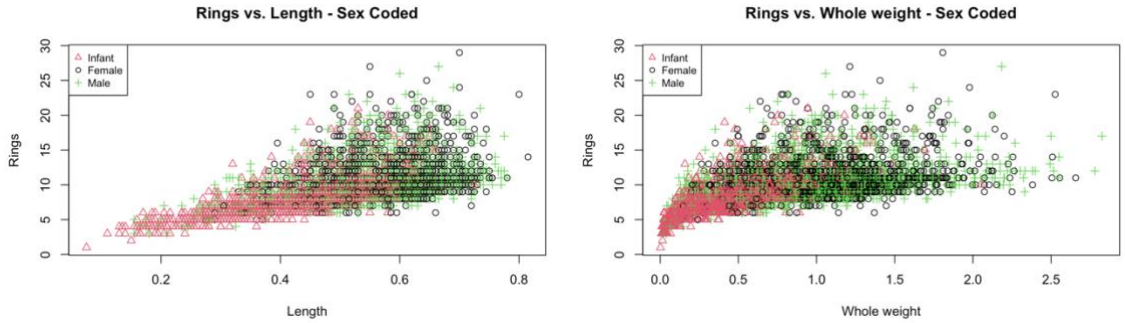

Figure 3. Sex coded scatters

From Figure 2, we can see that there is a significant linear relationship between whole weight, shucked weight, viscera weight, and shell weight, as well as between length, diameter, and height. There is a polynomial relationship between weight and length. As we take sex, length, diameter, height, whole weight, shucked weight, viscera weight, shell weight as X_1, X_2,…, X_8, and age as Y, there is no significant nonlinearity between Y and X, but there is a high degree of multicollinearity between X. Therefore, forward stepwise screening of variables is considered. From Figure 3, we can see that sex at maturity has little effect on the relationship between age and length-weight.

2.2 Model Selection in Multiple Regression

As our goal is to have a best predicting model, we first split our data into training set

and test set, and build models on the traning set. We conduct various stepwise regression searches with the first-order model as the full model. Whether using AIC or BIC criterion, forward selection, backward elimination, and forward or backward stepwise result in the same model in which length is dropped:

$$Y = \begin{cases} 5.215 + 7.384X_3 + 23.095X_4 + 8.483X_5 - 18.813X_6 - 10.008X_7 + 7.507X_8, \ X_1 = Female \\ 5.202 + 7.384X_3 + 23.095X_4 + 8.483X_5 - 18.813X_6 - 10.008X_7 + 7.507X_8, \ X_1 = Male \\ 4.401 + 7.384X_3 + 23.095X_4 + 8.483X_5 - 18.813X_6 - 10.008X_7 + 7.507X_8, \ X_1 = Infant \end{cases} \quad (1)$$

Table 2. VIF of the best first-order model

	X_1	X_3	X_4	X_5	X_6	X_7	X_8
VIF	1.519	9.424	6.618	97.858	26.000	16.504	20.025

From the VIF value, we can see that the model shown in formula (1) has severe multicollinearity on X_5 , whole weight, which corresponds to Figure 2. The interpretation has little difference between Male and Female, and the p-value(0.768) of Male is high, which means there is no statistically significant difference between Female and Male in Sex dummy, corresponding to Figure 3. So the next step is to remove the X variables with high VIF values and set Female and Male as the same category in Sex. Also, Figure 3 illustrates that Y needs to be transformed to $Y^{-1/4}$.

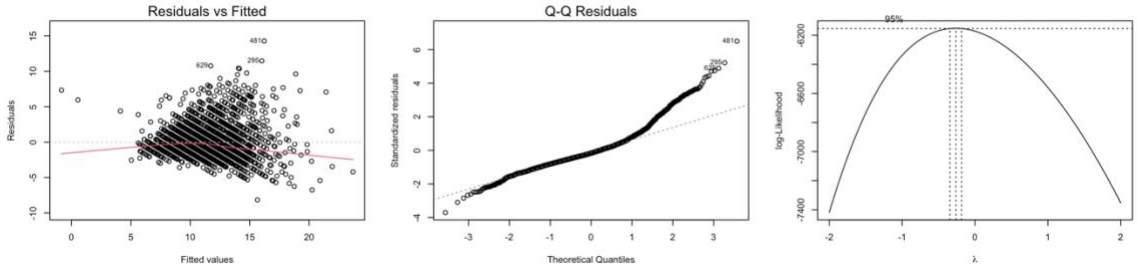

Figure 3. Residuals plot, Q-Q plot, and Box-Cox result of the first order model

After Y transformation and Sex category combining, we get Model 1 from Stepwise Regression Procedures:

$$\frac{1}{\sqrt[4]{Y}} = \begin{cases} 0.656 - 0.200X_3 - 0.292X_4 - 0.034X_5 + 0.168X_6 - 0.077X_8, \ X_1 = Female \text{ and } Male \\ 0.666 - 0.200X_3 - 0.292X_4 - 0.034X_5 + 0.168X_6 - 0.077X_8, \ X_1 = Infant \end{cases} \quad (2)$$

After deleting X variables with high VIF, we get Model 2:

$$\frac{1}{\sqrt[4]{Y}} = \begin{cases} 0.658 - 0.205X_3 - 0.307X_4 + 0.126X_6 - 0.127X_8, \ X_1 = Female \text{ and } Male \\ 0.669 - 0.205X_3 - 0.307X_4 + 0.126X_6 - 0.127X_8, \ X_1 = Infant \end{cases} \quad (3)$$

Then we take a look of the model plots.

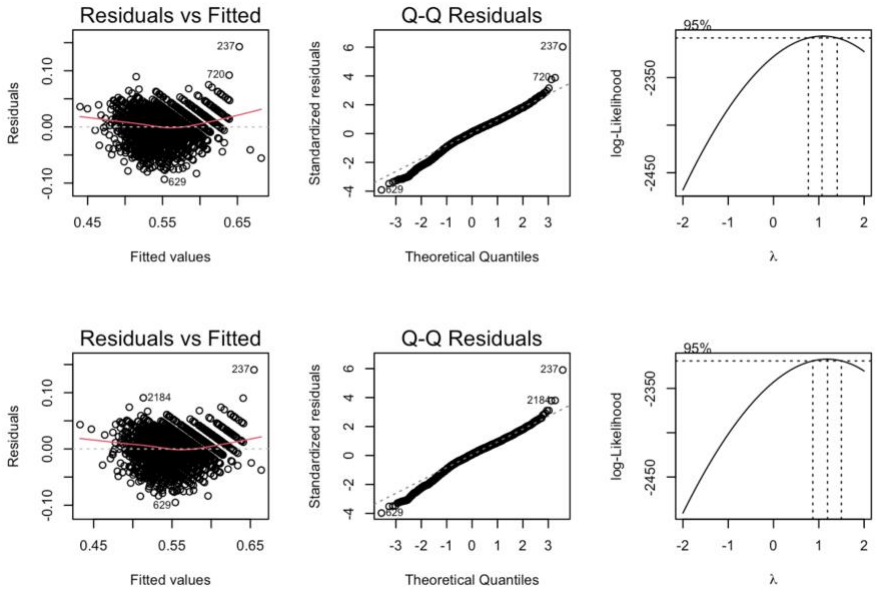

Figure 4. Residuals plot, Q-Q plot, and Box-Cox result of Model 1and 2

From the plots, the normality hypothesis is satisfied now, but as the residuals plots still shows a curvilinear trends, indicating that the model may have a nonlinear relationship. We do not consider interaction terms in our model because the VIF of interaction terms is extremely large. So we try to add quadratic and cubic terms of Diameter, Height, Shucked Weight and Shell Weight to the model.

When the criterion is AIC, we get Model 3 from Forward Stepwise Procedures:

$$\frac{1}{\sqrt[4]{Y}} = \begin{cases} 4.764 + 43.093X_4 - 81.8X_4^2 - 18.897X_6 + 7.646X_6^2 + 47.575X_8 - 55.488X_8^2 + 31.014X_8^3, \\ X_1 = Female\ and\ Male \\ 4.049 + 43.093X_4 - 81.8X_4^2 - 18.897X_6 + 7.646X_6^2 + 47.575X_8 - 55.488X_8^2 + 31.014X_8^3, \\ X_1 = Infant \end{cases} \quad (4)$$

When the criterion is BIC, we get Model 4 from Forward Stepwise Procedures:

$$\frac{1}{\sqrt[4]{Y}} = \begin{cases} 5.835 + 20.350X_4 - 81.8X_4^2 - 18.740X_6 + 7.425X_6^2 + 52.040X_8 - 67.872X_8^2 + 39.615X_8^3, \\ X_1 = Female\ and\ Male \\ 5.120 + 20.350X_4 - 81.8X_4^2 - 18.740X_6 + 7.425X_6^2 + 52.040X_8 - 67.872X_8^2 + 39.615X_8^3, \\ X_1 = Infant \end{cases} \quad (4)$$

3 Model Prediction

We now have 4 models fitted the training set and want to know which model gives the best prediction. We predict the age of abalone with 4 models on both training set and test set to see if any of them is overfitted. From Table 3, as all of the models have similar MSE for training sets and test sets, they are not overfitted.

Table 3. Model Validation – MSE of Training set and Test set

	Model 1	**Model 2**	**Model 3**	**Model 4**
Train set MSE	128.5656	128.5643	4.850329	4.862254
Test set MSE	129.8486	129.8505	4.850329	4.553147

Table 4. Model prediction comparison

	Model 1	**Model 2**	**Model 3**	**Model 4**
MSE	129.8486	128.5643	4.5445	4.5531
RMSE	11.3951	11.3952	2.1318	2.1338
MAE	10.9129	10.9130	1.5493	1.5498
R2	-11.2288	-11.2289	0.5720	0.5712

From Table 4, we can see the Model 3 have the smallest MSE, RMSE and MAE, and the largest R2, so we say Model 3 is the best predicting model.

4 Conclusions and Discussion

From our analysis, the model 3 is the best predicting model with the smallest MSE, RMSE and MAE, and the largest R2 on the test set:

$$\frac{1}{\sqrt[4]{Y}} = \begin{cases} 4.764 + 43.093X_4 - 81.8X_4^2 - 18.897X_6 + 7.646X_6^2 + 47.575X_8 - 55.488X_8^2 + 31.014X_8^3, \\ X_1 = Female\ and\ Male \\ 4.049 + 43.093X_4 - 81.8X_4^2 - 18.897X_6 + 7.646X_6^2 + 47.575X_8 - 55.488X_8^2 + 31.014X_8^3, \\ X_1 = Infant \end{cases} \quad (4)$$

where Y is the age of abalone, X_1 is the sex, X_4 is the height, X_6 is the shucked weight, X_8 is the shell weight. This model is not overfitted and suggests that incorporating higher-order terms or interaction terms in the model may improve predictive accuracy.

The results are specific to the abalone dataset used and may be limited to datasets with similar features. While the model could potentially apply to other datasets with similar structures, its effectiveness may vary if the characteristics of other populations differ significantly. Another limitation is the potential overfitting observed in our complex models with higher order terms. Although it performed well on our dataset, its complexity might lead to poorer generalization on new data sets. To address these limitations, future work could include external validation on an independent abalone dataset to assess the model's generalizability. Additionally, employing regularization techniques, such as Lasso or Ridge regression, could help reduce overfitting and improve the model's robustness. Collecting more comprehensive data with a larger and more diverse sample could also enhance the generalizability of our findings.